Multi-structured Data Integration and Analysis for Decision Making

***Abstract* — Intelligent decision making highly relies on ability to combine data from different sources and generate appropriate insights. Multi-structured data analytics holds special significance for organizations, where different types of data are produced by different sources and applications, and assimilating information from this diverse collection poses huge challenges. Data collection can contain information in the form of numbers, aggregates, time-series, reports, proposals, logs, communication records etc. Insight generation in such a scenario requires novel methodologies to judiciously link information from multiple sources and reason with them. This paper focuses on developing methodologies for intelligent integration of multi-structured data. The emphasis is on intelligent methods of linking information amongst data from unstructured sources and to data that is obtained from structured sources. Different types of correlations are identified which include associative correlations as well as possible causal correlations. We propose methodologies to use important events from text documents and assess their impacts on quantitative data.**

Keywords—events; multi-structured data analytics; event extraction; data integration;

I. Introduction

Enterprise decision makers accept that the first step to intelligent decision making is to opt for data integration from multiple sources both within and from outside the organization. Enterprises have realized the necessity of breaking their fire-walls in order to understand their customers better. Social media monitoring has become an integral part of corporate decision making. Though the big data storage and access technologies have provided them with the required platforms to harness the data, integrating relevant information from diverse sources and utilizing it all for decision making purposes is still a huge challenge. To begin with large portions of this data is unstructured, diverse in nature and most often generated in an absolutely uncontrolled fashion. Thus there are multiple challenges that need to be addressed in order to derive meaningful business insights from this data. Some of the well-known challenges typical of handling big data, namely integration of large volumes of data coming at a very high rate from disparate sources is being addressed by multiple vendors through improved technology. However, a challenge that is still not well-document and hence has received little attention is lack of clarity on the nature of insights that can be derived from integrated multi-structured data. Lack of methods and analytical techniques to derive such insights and use them for decision making is obviously the next challenge.

Multi-structured data analysis holds special significance for organizations since they are measured by performance indices that are numerical quantities or numerical-valued time-series data like monthly sales figures, market share, customer satisfaction indices, stock-values etc. However, analyses interpretation of these data demand active integration with information obtained from unstructured data sources like market News and reports, financial and political News of a region, competitor reports, customer opinions and reviews, customer communications about product or service related problems, stake-holder communications like vendor communications, dealer-reports, service-agency reports etc. in order to utilize the information effectively for decision-making. While numerical data can only reveal movements and trends, the causes for these can be often ascertained using information extracted from unstructured data. In recent times contribution of social media content to this process has gone up significantly. Since more than 2 billion people share their experiences on social media, it is becoming easier to identify trends. Smart techniques can help organizations analyze large volumes of relevant information to understand early trends. Early trends help in preventing disasters as well as grab opportunities for competitive edge. Consumer product manufacturers can make use of this data to dynamically predict their demands. The data can also be used to identify customer wish-lists and therefore use it for designing new products. The possibilities are end-less. While some of these things are utilized in decision-making today - it is only happening through isolated searches and off-line integration of results. We don't claim to have solved the problem but present some promising initiatives towards reaching there in not so distant a future.

In this paper, we present a framework that aids in acquisition and integration of data from multiple sources specifically to enrich business intelligence. The paper also presents methods for integrating structured enterprise data like sales figures, market-share or stock-index reports with Open Source News articles and consumer-generated social-media content. The aim of the framework is to provide analysts with an integrated environment where data is effortlessly pulled in from multiple sources in order to provide deeper insights about related business events. We emphasize that decision-making is still a human-expert activity. The role of the present framework is restricted to help in the decision-making process by presenting a wide-range of relevant information. The algorithms and methods are therefore focused on judging relevance of different pieces of information to a business event and presenting them to the user in a comprehensible manner.

The contributions of this paper are as follows:

(i). It presents methodologies to mine significant business events from domain-specific News documents. Significance of an event is defined as a function of its occurrence pattern in large relevant collections.

(ii). It proposes novel methods to associate significant Business News events to social-media content. Information gathered from social media and public forums can capture public sentiment reactions to a business event and can therefore provide crucial insight about business performance.

(iii). It provides a framework to relate structured data to relevant unstructured content through contemporary significant business events. The framework is navigational and allows analysts to scour through all related pieces of information in order to verify and consolidate.

Results from different real data sources are presented to establish the effectiveness of the proposed methodologies and also show the applicability across different domains and different types of data analyses tasks. We believe that this work will pave the way for automated analysis in future wherein the impact of the events on the numbers can be assessed from the associations observed over large collections.

The rest of this paper is organized as follows. Section 2 presents a survey of related work. Section 3 presents the framework for multi-structured data analytics. Section 4 presents the process of detecting significant events from structured data. Section 5 presents methodology for extracting business events from news. Section 6 explains social media information extraction. Section 7 presents event-driven multi-structured data integration process. Section 5 presents experiments and results from automobile domains using about 24 months of relevant News and social media data.

II. RELATED WORK

Multi-structured data analytics has received prominence with the interest in big-data which aims to integrate data from multiple sources. Early attempts at integration analytics were presented in [9, 10] which described applications that automatically associated unstructured content generated through customer interactions like emails and transcribed phone calls with customer and account profiles stored in an existing data warehouse based on pattern-driven Named-entity matching. The insights derived through these associations helped organizations to understand the changing needs of the customers in a timely manner, and also use them to introduce new products and services, as well as to improve and personalize the interactions. However these applications did not consider linking open source data with enterprise data.

One of the popular sub-tasks of information extraction (IE) from open source documents is to extract events from large volumes of unstructured text like News collections [3] or social-media content like Tweets. REES [1] reported a large-scale event extraction system that extracted 61 pre-specified types of events under different categories like crime-events, business-events, financial-events, political-events etc. using a declarative, lexicon-driven approach where each lexicon entry is defined for a specific type of event. [2] presents a real-time and multilingual news event extraction system to extract violent and natural disaster events from online news using shallow NLP techniques and cluster them. While the earlier works focused on discovering pre-specified events, methods to dynamic event discovery mechanisms were reported in [4]. This was based on dynamic discovery of relationships among co-bursting entities along with underlying global and local time constraints. In [5], a model was proposed that jointly performs event extraction and segmentation. The model exploits distribution of the field values in text to identify event boundaries. It employs sentence-level event field consistency constraints as well as document-level event consistency constraints to identify and segment events correctly. Joint inference is facilitated by the use of CRFs for each sub-task.

With growing focus on Twitter as a source of News, event extraction from tweets has also gained popularity. CRF based event extraction methods were reported in [6]. [6] presented methods for categorizing important upcoming events. In [7], methods were proposed to link related events which could depict evolution of an event.

It may be observed that none of the work considered business event extraction and correlation of business events with significant events happening within an enterprise. Since most business events are not likely to get world-wide coverage, these are unlikely to be discovered through statistical techniques. The uniqueness of the proposed work is that while its objective is to find all possible business events from a given collection of News articles, attempt is also made to determine the significance of different classes of events within a time-period.

III. MULTI-STRUCTURED DATA ANALYTICS - A FRAME-WORK

The core functionality of the framework can be stated as follows:

(a). Correlate observations on business data which is in the form of time-series with relevant News events. The underlying reasoning framework supports intelligent correlation mechanisms that use a combination of available meta-data like time-of-occurrence, source etc. along with more complex concepts extracted from News articles through the use of natural language processing and text mining techniques.

(b). For each association generated through correlation of structured data and News articles, retrieve relevant information from related social-media content stored in the repository.

(c). Present all information together to an analyst with facilities to further drill down and explore additional material.

Fig.1 illustrates an overview of the proposed framework for acquisition, integration and analysis of multi-structured data from multitude of sources. This is an extension of the framework presented in [7], which stated some basic ideas about how structured and unstructured data can be integrated.

A. Data Acquisition

Structured data required for the purpose can be collected from multiple sources. The framework has built-in functionality to collect stock-market data from Yahoo! Finance. It also supports ingestion of data from associated Data-bases. In a case-study presented in this paper, we show the use of sales volumes and ranks of competing products, which are collected from open-source. Structured data is also time-stamped and has associated meta-data like organization names, product names, source of information etc.

The framework supports acquisition and assimilation of domain-specific text documents from multiple sources like News sites, analyst web-sites, enterprise News sites etc. Site-specific crawlers are integrated for the purpose. News articles can also be collected using RSS feeds.

Another source of unstructured text content is social-media. The framework supports integration of dedicated crawlers for acquisition of consumer-generated content from organization-specific dedicated web-sites. It also supports collection of tweets containing specified key-words or user-ids using Twitter's streaming API.

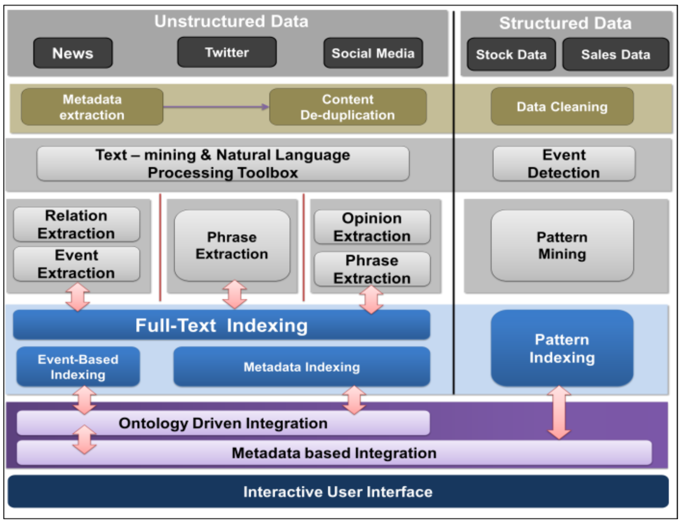


Figure 1 Architecture

All gathered content is uniquely identified, time-stamped and stored along with additional meta-data elements like author name, source-name, category of document, ratings if any, specific named-entities like product names etc. if they are a part of information documentation on site and any other information that the site may provide. The unstructured content is subjected to a series of information-extraction methods based on Natural Language Processing (NLP) techniques and text-mining methods for extraction and compilation of embedded information for integrated analysis.

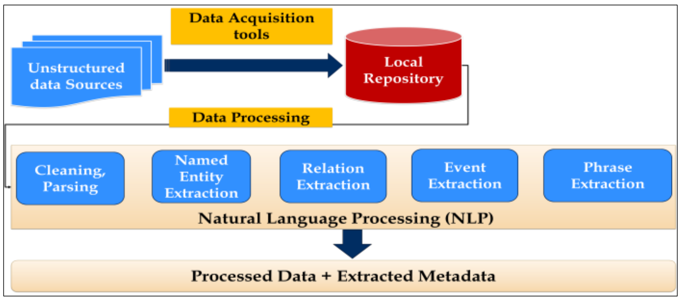


Figure 2 Unstructured data processing pipeline

B. *Unstructured Data Processing Pipe-line*

The content processing pipe-line for unstructured text includes a set of pre-processing, cleaning, Natural Language Processing and text mining tasks. Figure 2 presents methods that are applied sequentially to all News articles, tweets and consumer-generated feedback that are collected through crawling. Various tools and technologies that are already available or details of which have been published earlier, are deployed to extract relevant information from text. The extracted information is stored along with information about the source. The activities involved in each step are mentioned briefly.

(i). Pre-processing and Cleaning - Crawled content from a web-site is broken into individual pieces and tagged by unique identifiers and associated meta-data. Since consumer-generated content is very noisy, this step also includes tasks like sentence identification, special-character removal etc.

(ii). Parsing - Stanford Parser is employed to generate Parts-of-Speech (POS) tags for words and also parse the content.

(iii). Phrase extraction and normalization - It is observed that Stanford parser misses a lot of phrases when the underlying content is noisy, since the sentences are often grammatically incorrect. Hence, the phrase set is enhanced using techniques described in [11]. Using the extracted phrases as seeds, phrases similar to the seeds, are extracted from the texts. The enhanced set contains phrases which may be semantically similar though not identical to the original ones and also those that could not be identified due to grammatical errors like misplaced punctuations or incorrect use of prepositions etc.

(iv). Named-entity recognition - Named-entity recognition (NER) seeks to locate and classify elements within text into pre-defined categories such as names of persons, organizations, locations, products etc. Named Entity Recognizers can also recognize and extract special expressions like times, quantities, monetary values, percentages, etc. and their corresponding values. We have employed the Stanford Named Entity Recognizer (Stanford NER) for our work. Named entities in subject or object of sentences provide information about topics of discussion. Hence they play an important role in integrating unstructured information from multiple sources. However, since all objects of interest are not named entities, the set of named entities are further extended to include enhanced set of noun-phrases for the purpose of data integration.

C. *Overview of Data Integration Process*

The first level of integration occurs through the meta-data. One of the easiest ways to link multi-structured data is through time-stamps. However, integration for decision-making requires context-based associations, which need deeper analysis of content. As presented in [7], the present framework supports this through an event-based analytical and reasoning system.

Events are a popular and succinct way to represent information reported in News. An event is defined as something that happens at some specific time and place [8]. As pointed out in section 2, extraction and compilation of News events is an active research area. In this work, the focus is not on generic event extraction, but on learning to extract and characterize significant business-events from focused News collections. The event-extraction and characterization methods are presented in the next section. These methods are different from the basic event extraction methods presented with the earlier version of the framework in [7].

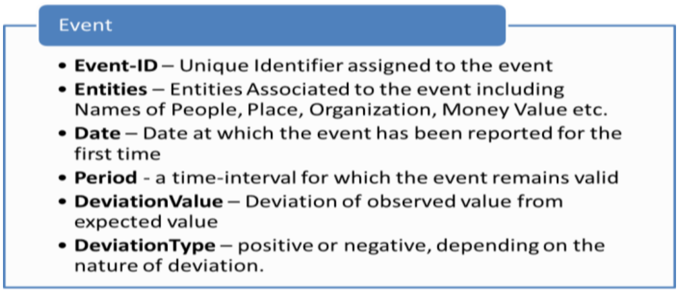
The current framework also mines events from the structured data using the same methods presented in [7]. Structured data are time-series values, representing performance factors of an enterprise observed at discrete time instants. Events in the structured data space are therefore modeled as significant statistical deviations from running averages. Deviations from running averages can detect events like sudden surge or fall in sales volume, rank or market-share of a product or an organization as a whole. It can also detect and characterize significant stock-market movements observed over a short-term. Long term stock trend movements require more complex event detection techniques. The present framework does not have built-in methods for detecting long-term trends.

Social-media content helps the analyst understand whether there were early social-media signals related to the event. The framework thereby supports mechanisms for learning significant contextual associations from large focused multi-structured data-sets to aid the decision-making process. The ultimate aim is to equip an organization to detect early signals from the Web and social media for predictive reasoning.

The framework also supports the use of domain ontology to aid integration of data and information extracted from multiple sources.

IV. DETECTING SIGNIFICANT EVENTS IN STRUCTURED DATA

The framework supports analysis of enterprise structured data that are represented as Time-series [7]. Significant events are therefore detected as deviations in observed behavior of a measurable variable from its expected behavior. Simple events can be defined in terms of rise or fall of the value or as deviations from expected value that has been predicted by a model. Expectations are computed as rolling averages. Deviations are characterized using pre-defined thresholds.



Event

- **Event-ID** – Unique Identifier assigned to the event
- **Entities** – Entities Associated to the event including Names of People, Place, Organization, Money Value etc.
- **Date** – Date at which the event has been reported for the first time
- **Period** - a time-interval for which the event remains valid
- **DeviationValue** – Deviation of observed value from expected value
- **DeviationType** – positive or negative, depending on the nature of deviation.

Figure 3 Structured data Event representation

More complex events can be defined as functions of state-changes over time or as functions of multiple time-series. For example, while defining a stock-market deviation event for a company we have made use of stock values of the sector as a whole tracked over a defined time-period, rather than looking at isolated values.

Since each time-series can be associated with a set of specific entities, a structured time-series event is represented as shown in Figure 3. Figure 4 shows Toyota's stock graph with deviation events marked with green triangle.

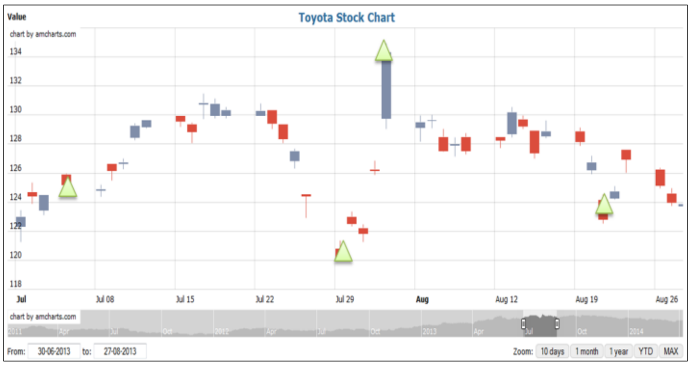


Figure 3 Toyota's stock graph with deviation events

V. EXTRACTING BUSINESS EVENTS FROM NEWS

Since events by definition report "happenings" or some activity, event extraction focuses on analysis of "verbs" in sentences. Event characterization uses verbs, their synonyms and categories along with associated Named Entities.

Ollie is a pattern-based relation extractor that analyzes sentences to identify embedded triplets representing (SUBJECT - PREDICATE - OBJECT). While the subject and object contain names of entities that take part in the event, the predicate contains information about the nature of the activity reported. Ollie can capture relation phrases expressed by verb-noun combinations after eliminating incoherent and uninformative extractions.

It works with the tree-like (graph with only small cycles) representation using Stanford's compression of the dependencies. The relation extractor produces outputs of the form (SUBJECT – PREDICATE – OBJECT) where the predicate is as follows:

V | VP| VW*P
V = verb particle? Adv?
W = (noun | adj | adv | pron | det)
P = (prep | particle | inf. marker)

It was observed that the quality of predicates extracted by Ollie was noisy and not enough to be identified as an activity. This is partially due to the complex structure of the sentences and partially due to the noise inflicted by the content extraction process which works with arbitrarily designed Web pages. Hence, we additionally identify the key verb within the predicate based on the POS analysis and use this for event-action detection.

Obviously, every verb does not represent a significant business activity. Also significant business events are usually reported in multiple articles. Often these repetitions are spaced out in time. For example, a major recall for a popular automobile model is covered by different News agencies and may also be referred to in a quarterly or an annual report. It may be pointed out that differences may be observed in the details like exact numbers or money values etc. even when the same event is reported in multiple sources. These issues need to be taken care of while linking News events to structured data events to aid contextual interpretation of numerical data.

Based on the above observations we now re-formulate the task of business event detection and characterization in terms of three sub-problems:

1. Identification of domain verbs that represent significant event categories for a given domain.
2. Grouping similar event instances from across the repository.
3. Estimate the importance of an event based on its occurrence across different articles within a repository.

A. *Identifying Significant Domain Verbs*

Since verbs are key identifiers for events, significance of a verb is assessed based on its occurrence pattern over a time period. It is observed that some verbs have uniform and regular occurrence patterns with very high frequencies. Some verbs on the other hand exhibit sudden spikes at irregular intervals. Yet another set of verbs show spikes at regular intervals. Frequencies of these verbs may or may not show large variations. Though the verbs may differ with the domains and collections, these behavioral aspects are observed across different focused collections. These observations led to the use of entropy as a measure for significance.

Given that a document repository is observed for time period **T** divided into **t** intervals, the significance of a verb is given by the average information content, measured by entropy, over the time period. Significance of a verb **v** is denoted by $\boldsymbol{\sigma}(\mathbf{v})$ and computed as follows:

$$\sigma(v) = \left(\frac{1}{t}\right) * \sum_{i} E_i\,(v),$$

$where\ E_i(v) = -p_i * \ln(p_i)\,, i = 1..t$ and p_i is the probability of verb v occurring in the documents time-stamped by the ith time interval.

Table 1 shows the top 25 verbs identified for two different News repositories. The first collection comprised Stock Market News collected for 6 major IT companies of India. The second collection contained Market News for major Automobile companies of US.

TABLE 1 SIGNIFICANT DOMAIN VERBS TO IDENTIFY IMPORTANT EVENTS

Domain	Significant Verbs
Stock Market News for Indian IT Companies	say, report, announce, provide, trade, win, take, see, give, rise, close, continue, make, grow, post, deliver, add, work, acquire, offer, hold, get, include, gain, fall
Automobile Market News from US	say, sell, make, include, take, offer, come, get, announce, build, use, go, plan, continue, recall, tell, provide, add, see, start, report, give, look, expect, work

Analysts can restrict themselves to work with top N events to perform business analysis. Choosing a very small value for N may cause an analyst to miss out on significant but rare business events. Too big a value for N may introduce noise.

B. *Grouping Similar Event Instances*

In order to group similar events, we employed Min-hash algorithm proposed in [20] to first group the similar sentences together and then associated a structured representation of an event to the group. To group similar sentences across the repository, the word-vector representation of the sentence is used after stop-word removal. The following steps are then applied to group similar sentences:

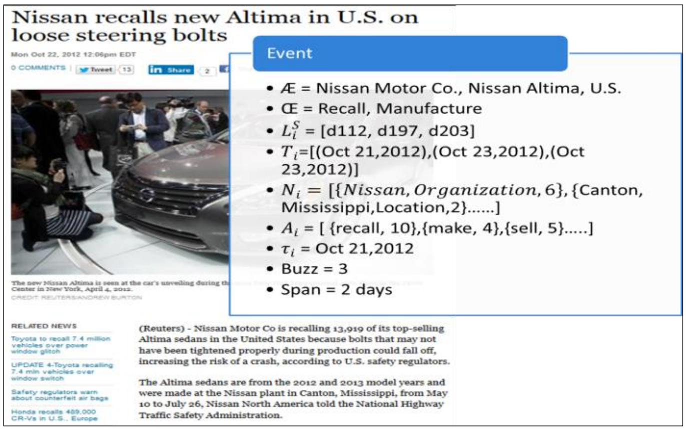
1. The word-vector of each sentence is converted to another vector using 16 Min-hash functions.
2. Projections are taken of this vector on 4 dimensional hyper-planes.
3. Events with same projection form a neighborhood.
4. A set of sentences which lie in same neighborhood and have similarity value more than 0.75 on Jaccard Similarity metric are clubbed to a group.
5. The group is assumed to represent an event with possibly multiple instances of it observed across the repository. Each group is assigned a unique identifier termed as the Event-ID and has an associated structured representation defined later.
6. Named-Entities and Verb grouping – Named entity set for an event is constructed by combining all named entities associated to the events belonging to the group. Named entities of a specific type within this set are resolved on the basis of string matching and containment. The string with the longest length is chosen as the representative for a set of entities that are resolved to be same. The resulting list of all distinct named entities along with their types is associated as event metadata. Verbs, which are maintained in their root forms are also combined and stored as event meta-data. The entity and verb sets are further used for event naming.
7. Event-naming - Event naming consists of tagging an event with possible actors and a named activity.
    a. Activity Naming - The most frequent verb from the verb list above is used to name the event. It may be noted that activity names can be multi-values in case of multiple verbs having same maximum frequency. Further work is in progress to use synonyms and verb class-names.
    b. Actor Naming - The named entities from the named entity list is used to identify actors for an event. Only those entities that appear within subject and object chosen as actors for the event. This step helps in ensuring that mostly named entities of type people, organization and location are associated as actors, thereby eliminating named entities of type money-value, time etc. from actor list.

Let Ê be the set of all unique events extracted from a repository R containing a total of **M** documents. Let the cardinality of Ê be **N**. As remarked earlier, each element of Ê is representative of a set of event instances, where the instances of a set can be linked back to sentences in source documents.

Each event $E_i \in \hat{E}$ is assigned a unique Event-ID and contains the following derived components in its structured representation:

1. Each event E_i is associated with a pair of lists <Æ> and <Œ> , where Æ are the set of identified actors and Œ are the set of identified activity names.
2. L_i^S = List of source documents for E_i
3. T_i= Time-line is an ordered set of time-stamps, with possible duplication constructed from the publish-times of the source documents.
4. $N_i = \langle \text{Entity} - \text{ID}|\text{Type}|\text{Frequency} \rangle$ is a list of tuples recording Entity occurrences in the group along with their types and frequencies.
5. $A_i = \langle Verb|Frequency \rangle$ is a list of verbs extracted from the predicates along with their total frequencies in the group
6. τ_i = First-Report-Time of event = earliest time-stamp in T_i.
7. Buzz = cardinality of L_i^S
8. Span = Difference between the most recent time-stamp and the earliest time-stamp in T_i.

Figure 4 shows a News article along with one event of activity type Recall extracted from this source, with all its details. The details show that this event was first reported in media on 21 Oct, 2012 and occurred in multiple News articles till 23 Oct, 2012. The actors associated with this event are Nissan Motor Co., Nissan Altima etc.



Nissan recalls new Altima in U.S. on loose steering bolts

Mon Oct 22, 2012 12:06pm EDT

The new Nissan Altima is seen at the car's unveiling during th
Center in New York, April 4, 2012.
CREDIT: REUTERS/ANDREW BURTON

Event

- Æ = Nissan Motor Co., Nissan Altima, U.S.
- Œ = Recall, Manufacture
- L_i^S = [d112, d197, d203]
- T_i=[(Oct 21,2012),(Oct 23,2012),(Oct 23,2012)]
- $N_i = [\{Nissan, Organization, 6\}$,{Canton, Mississippi,Location,2}......]
- A_i = [{recall, 10},{make, 4},{sell, 5}.....]
- τ_i = Oct 21,2012
- Buzz = 3
- Span = 2 days

RELATED NEWS

Toyota to recall 7.4 million vehicles over power window glitch

UPDATE 4-Toyota recalling 7.4 mln vehicles over window switch

Safety regulators warn about counterfeit air bags

Honda recalls 489,000 CR-Vs in U.S., Europe

(Reuters) - Nissan Motor Co is recalling 13,919 of its top-selling Altima sedans in the United States because bolts that may not have been tightened properly during production could fall off, increasing the risk of a crash, according to U.S. safety regulators.

The Altima sedans are from the 2012 and 2013 model years and were made at the Nissan plant in Canton, Mississippi, from May 10 to July 26, Nissan North America told the National Highway Traffic Safety Administration.

Figure 4 News article with extracted event

C. *Estimating the Impact of an Event*

Events do not follow a uniform or well-defined probability distribution throughout the collection. While some events have a more regular occurrence pattern, others have widely varying occurrence patterns. For example, since all industries publish quarterly sales report, it is likely that the occurrence of "sell" verb and thereby more events of the corresponding type will be observed in the collection in specific months of the year. However, events like "recall" or "acquisition" are highly irregular in nature and cannot be expected to follow a well-defined distribution over time. Event occurrences such as these are also more likely to affect external events or be affected by them. While some event mentions in text can be identified as cause for an external event observed in numerical data, a pattern in the numerical data may also give rise to certain event mentions in a collection as its effect. For example, while an acquisition may cause the stock-value of a company to rise, fall in sales of a particular model may trigger the company to reduce price or offer discounts to the customer, which may be reflected as announcements or offers in the collection. In this section we propose methods to identify nature of the verbs. This in turn can help in distinguishing between routine events and irregular events.

We now state methods for detecting significant events for a given entity E, for a time-instant P. This is done using linear regression analysis over past event observations. Regression analysis is a statistical technique for estimating the relationships among variables. In our case, we analyze the history of occurrence probabilities of events over specified time-intervals prior to P.

The intervals observed for real business data are usually weeks or months, but could be quarters or years also. For a given repository, the total time period T for the document collection is also split identical time-periods denoted by $\{t_1, t_2, \ldots, t_k\}$. The choice of length of t_i is dictated by the business data.

Let $C = \{c_1, c_2, \ldots, c_N\}$ be N types of event categories, associated to top N significant domain verbs. For each c_i in C, we can associate a series $\{(t_i, y_i)\}$, where y_i denotes the observed frequency of event of type c_i in time-period t_i.

A regression line is defined by $y = a_0 + a_1 t$, with regression coefficients calculated as follows:

$$a_1 = \frac{m \sum_i t_i y_i - \sum_i t_i \sum_i y_i}{m \sum_i t_i^2 - (\sum_i t_i)^2}$$

$$a_0 = \frac{\sum_i y_i}{m} - \frac{a_1 \sum_i t_i}{m}$$

Using this regression line we predict frequency of the event c_i as $\widehat{y_{m+1}}$ for the time period t_{m+1}. If the observed frequency of the event for t_{m+1} is y_{m+1}, then residual is $y_{m+1} - \widehat{y_{m+1}}$.

An event is termed as significant, if the residual is very high; indicating that the documents published within time period t_{m+1} has unusual high frequency of event c_i.

Impact of an event on actual enterprise data is now estimated based on its occurrence across time-periods including future references also. The intuition behind this step is simply to use the fact that an event which is judged as impactful by the industry analysts would be published more across different sources. Thus in a way, the framework is exploiting the expertise of human analysts, which manifests as frequent repetitions in a collection. It can also be equated to the concept of social-media buzz, where the more popular content becomes, the more likely it is to be shared and therefore impact social-media users.

VI. INFORMATION EXTRACTION FROM SOCIAL MEDIA DATA

Business analysts are unanimous about the high potential of social-media content in deriving business intelligence. Numerous incidents in recent history have shown that social sentiments can play both positive and negative roles in driving business results. All these have proved beyond doubt the need to integrate social-media content into traditional analytics platforms. The proposed framework not only supports integration of relevant social-media content with enterprise data and open-source News data, but also proposes methods for exploiting the information within an analytical framework.

Two types of social-media content are considered in this framework

(i). Tweets - Twitter is an important source of people opinion. It captures instant reactions of people to all kinds of events and has been proved to have mass influence. Tweets may contain explicit reactions of people towards an entity. However, often people's interest in an entity is elicited by the fact that they re-tweet a News item about an entity without generating any new content. Though no new information is generated in the process, the rising trends are indicators of people's interest in the content. Thus the importance or impact of tweets can be captured from their duplication volumes or buzz counts.

(ii). Dedicated customer opinion websites - These sites also contain consumer-generated content. However these content have much more information including people's experiences, reactions and causes for reactions. The shelf-life of such content is also much higher. While tweets are useful to gauge consumer pulse at a point of time, this content is more suitable for deriving actionable intelligence. We will show later that this content can be used for predictive intelligence.

A. *Correlating News Articles and Tweets*

Though the potential uses of tweets are many in the context of business intelligence, the present framework uses tweets in a restricted context only. At present, tweets are only used to understand consumer reactions to News events. This is done through two steps:

(i). For a given business event extracted from News articles, retrieve *relevant* tweets from the storage.

(ii). Group the retrieved tweets. Each group is characterized by its frequent phrases and entities located within the group.

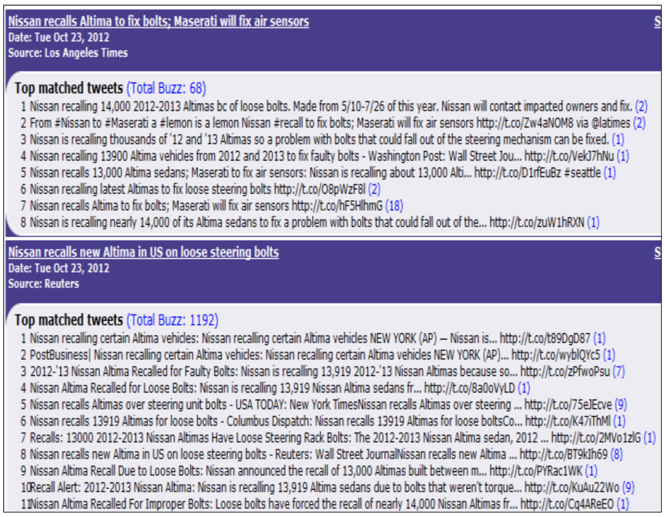
(iii). Return buzz for each group. Buzz is an indicator of consumer reaction.

Algorithm for retrieving matched tweets and grouping them - For a given business event ḣ which has been first reported at time ṫ and is associated with its source News articles denoted by set Ṅ, matching tweets are identified using the enhanced phrases extracted from Ṅ and tweets which were published within time (ṫ-δ_1, ṫ+ δ_2). The time-interval ensures temporal sanity. The matched tweets are clustered into groups based on their syntactic and semantic similarity using Bilingual Evaluation Understudy (BLEU) algorithm to detect similarities. The tweet retrieval algorithm is presented below.

Tweet Retrieval Algorithm

```
Input: Business Event ḣ first reported at time ṫ along with its source News articles Ṅ
Output: K tweets clustered into Groups along with top phrases of each group.
Step 1: Let P be the set of enhanced phrases extracted from Ṅ
Step 2: Let W be the set of words extracted from P
Step 3: Let C be the set of all tweets published within (ṫ-$\delta_1$, ṫ+ $\delta_2$) retrieved from local storage.
Step 4: For every tweet $t_p$ $\in$ C
        Step 4.1 If no group present, Make a new group containing Tweet $t_p$
        Else For tweet $t_q$ in every other group
            Step 4.1.1 If $t_p$ is substring of $t_q$ or $t_q$ is a
                substring of $t_p$ (this step identifies retweets)
                Step 4.1.1.1 add Tweet $t_p$ to group of $t_q$
                Else Calculate Score = BlueScore($t_p$,$t_q$)
                Step 4.1.1.2 If score >= 0.7 add $t_p$ to group
                of $t_q$
                Else Make a new group with tweet $t_p$
Step 5: For every group Tweet Group $T_G$ identified in step 4
        Step 5.1 Let $W_G$ represent all unique words in $T_G$
        Step 5.2 Let $s_g$ denote matching score for group $T_G$ = $|W_G \cap T_G|$
Step 6: Rank tweet groups based on Match_Score
Step 7: Output K = $\{T_{G,} W_G, s_g\}$.
```

Figure 5 shows a snapshot of two events along with their matched tweets.



Nissan recalls Altima to fix bolts; Maserati will fix air sensors
Date: Tue Oct 23, 2012
Source: Los Angeles Times

Top matched tweets (Total Buzz: 68)
1 Nissan recalling 14,000 2012-2013 Altimas bc of loose bolts. Made from 5/10-7/26 of this year. Nissan will contact impacted owners and fix. (2)
2 From #Nissan to #Maserati a #lemon is a lemon Nissan #recall to fix bolts; Maserati will fix air sensors http://t.co/Zw4aNOM8 via @latimes (2)
3 Nissan is recalling thousands of '12 and '13 Altimas so a problem with bolts that could fall out of the steering mechanism can be fixed. (1)
4 Nissan recalling 13900 Altima vehicles from 2012 and 2013 to fix faulty bolts - Washington Post: Wall Street Jou... http://t.co/VekJ7hNu (1)
5 Nissan recalls 13,000 Altima sedans; Maserati to fix air sensors: Nissan is recalling about 13,000 Alti... http://t.co/D1rfEuBz #seattle (1)
6 Nissan recalling latest Altimas to fix loose steering bolts http://t.co/O8pWzF8l (2)
7 Nissan recalls Altima to fix bolts; Maserati will fix air sensors http://t.co/hF5HlhmG (18)
8 Nissan is recalling nearly 14,000 of its Altima sedans to fix a problem with bolts that could fall out of the... http://t.co/zuW1hRXN (1)

Nissan recalls new Altima in US on loose steering bolts
Date: Tue Oct 23, 2012
Source: Reuters

Top matched tweets (Total Buzz: 1192)
1 Nissan recalling certain Altima vehicles: Nissan recalling certain Altima vehicles NEW YORK (AP) – Nissan is... http://t.co/t89DgD87 (1)
2 PostBusiness| Nissan recalling certain Altima vehicles: Nissan recalling certain Altima vehicles NEW YORK (AP)... http://t.co/wyblQYc5 (1)
3 2012-'13 Nissan Altima Recalled for Faulty Bolts: Nissan is recalling 13,919 2012-'13 Nissan Altimas because so... http://t.co/zPfwoPsu (7)
4 Nissan Altima Recalled for Loose Bolts: Nissan is recalling 13,919 Nissan Altima sedans fr... http://t.co/8a0oVyLD (1)
5 Nissan recalls Altimas over steering unit bolts - USA TODAY: New York TimesNissan recalls Altimas over steering ... http://t.co/75eJEcve (9)
6 Nissan recalls 13919 Altimas for loose bolts - Columbus Dispatch: Nissan recalls 13919 Altimas for loose boltsCo... http://t.co/K47iThMl (1)
7 Recalls: 13000 2012-2013 Nissan Altimas Have Loose Steering Rack Bolts: The 2012-2013 Nissan Altima sedan, 2012 ... http://t.co/2MVo1zlG (1)
8 Nissan recalls new Altima in US on loose steering bolts - Reuters: Wall Street JournalNissan recalls new Altima ... http://t.co/BT9kIh69 (8)
9 Nissan Altima Recall Due to Loose Bolts: Nissan announced the recall of 13,000 Altimas built between m... http://t.co/PYRac1WK (1)
10Recall Alert: 2012-2013 Nissan Altima: Nissan is recalling 13,919 Altima sedans due to bolts that weren't torque... http://t.co/KuAu22Wo (9)
11Nissan Altima Recalled For Improper Bolts: Loose bolts have forced the recall of nearly 14,000 Nissan Altimas fr... http://t.co/Cq4AReEO (1)

Figure 5 Nissan Altima recall events during Oct-2012 with Tweet groups

B. *Processing Customer Feedbacks*

Customer-opinion web-sites usually have provision for associating meta-data about the organization, product or service that are being discussed with the consumer-generated content. These details are also processed and stored as additional meta-data along with the content. For sites that support structured product rating, the rating is also stored along with the content.

Thus each feedback text has an associated product or entity name, time of feedback, rating and additional information like enhanced phrases that are extracted from it. As mentioned earlier, the framework assumes the existence of underlying domain ontology that contains lists of features and attributes for products or organizations under consideration. In order to quantify and structure the information content of feedbacks, appropriate domain ontology for the entity is instantiated.

Ontology labels at leaf nodes are assumed to represent feature and attribute names. These names are matched with enhanced phrases extracted from the feedback repository. A feedback is assumed to be about each ontology node for which a match is found in the text. Each feedback is indexed by all matching nodes contained in it across all its sentences.

In addition to the extraction of enhanced phrases, an additional text mining task that is undertaken for customer feedback processing is that of opinion extraction and sentiment tagging. Opinion mining and sentiment tagging are done using methods described in [11]. Opinions are essentially scored aggregated positive or negative values in association to an ontology feature or attribute that appears in the feedback text.

Aggregated opinion scores at leaf nodes are propagated along parent nodes. A feedback may also contain sentences that do not have any specific feature or attribute but contain sentiment expressions. The sentiment scores for these sentences are assigned to the root node. Sentiment scores aggregated at the root node, including those that are propagated from all its children are finally associated to feedback entity.

C. *Correlating Feedbacks to News and Tweets*

As mentioned in the earlier section, the key purpose of integrating customer feedbacks with sales data is to help the analysts understand whether there are any obvious correlations between consumer sentiments about a product and its sales data or between consumer perception of an organization with its market-share etc. It is a well-known fact that correlations do not indicate causation and blind statistical analysis can lead to meaningless correlations. The purpose of integration is to learn from human analysts what they consider as significant correlations. In future, we intend to extend the framework to learn from such human inputs to draw meaningful inferences.

In order to help such analysis, the proposed framework retrieves opinions and sentiments from relevant customer feedbacks pertaining to the task at hand. The relevance of a feedback to the task is determined by both the content and the period of analysis. Content-based relevance of a feedback to the task at hand determined as follows:

(i). If the focus of analysis is an entity, all feedbacks tagged with the entity name are judged as relevant with respect to content. The entity for which numerical data is analyzed is almost always available as meta-data.

(ii). Content based relevance of a feedback to a News article or tweets associated with it is decided by the overlapping sets of enhanced phrases and ontology labels.

A standard tag-cloud based visualization of text created with enhanced phrases is used to convey the feedback content to the analyzer. Opinion and sentiment scores are shown as normalized bar-charts. The phrases aid the analyst to have a view of problem reported by customers prior to an event. Figure 6 shows a snapshot of the system where top left displays the opinion graph, top right shows the phrases selected from customer feedbacks, bottom left shows phrases selected from news documents and bottom right shows customer opinion graph plotted for one of the selected phrases.

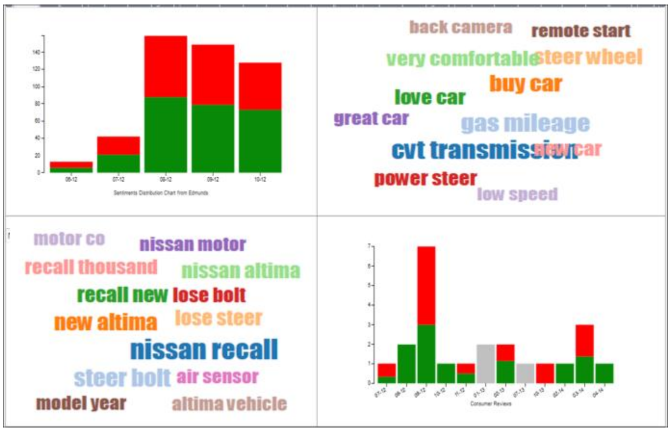


Figure 6 Customer feedback Sentiments and extracted phrases tag-cloud

VII. EVENT-DRIVEN MULTI-STRUCUTRED DATA INTEGRATION AND ANALYSIS

We now present the methods for event-oriented information retrieval to aid business intelligence analysts. As mentioned in the earlier sections - there are three types of data that have been processed, stored and indexed on different kinds of information that have been extracted from them. The analysis process starts with sequential retrieval and presentation of three types of data as mentioned below.

Step 1 For a chosen entity Ẽ, for a chosen time-period T, retrieve business data Ð as a discrete time-series.

Step 2 Discover significant business events on Ð as described in section IV. These events are ordered by decreasing order of impact. Let τ be a subset of T, such that each member of τ has an associated significant event.

Step 3 For each time t ∈ τ

Step 3.1 Retrieve all significant events E from News documents with the following properties and denote this set by Ῡa

(a). Each event in Ῡ has Ẽhas event in is set by gnificantentsub-section B of section V.

(b). For each event in in the First-Report-Time of event is within $(t \pm \partial)$, where ∂ is a pre-defined threshold usually varied from 1 to 3 days.

Step 3.2 For each event e_j ∈ Ῡ

Step 3.2.1 Retrieve all tweets in T that match with with e_j along with Buzz.

Step 3.2.2 Retrieve all stored social media content Ḿ discussing about Ẽ generated with time-stamp less than t.

Step 3.2.2.1 Show cumulative positive and negative sentiments about computed from Ḿ.

Step 3.2.3 Retrieve cumulative opinion scores for each attribute and feature of Ẽ as per ontology-based tagging of content in Ḿ.

VIII. EXPERIMENTS AND RESULTS

We present here some results from our data integration experiment conducted with data from the US Automobile sector. National Automobile Dealers Association publishes monthly sales data along with the volumes sold for top 15 models sold in US. It also publishes the market share for all companies listed in this sector. Our current repository has data collected from January 2012 onwards. News articles are collected through RSS feeds that come from all major News sources for the automobile sector. In this we present results using News articles ranging from January 2012 to March 2014. Customer feedbacks about relevant car models have been collected for the entire period from "Edmunds.com". However we could not collect tweets for the entire time-period due to resource constraints, and the collection only has tweets for the period October - 2012 to November 2012.

Figure 7 shows a snapshot that presents a complete statistics about the News sources and the share of 5 companies Ford, Toyota, Honda, Hyundai and Nissan in the News content over this period. Figure shows the rankings of a few top models of these 5 companies which are computed from the data gathered from NADA. The rank data shows some steep falls and rises, which serve as initial events of interest to an analyzer. In two separate case-studies we show how the data integration system helps analysts to systematically analyze such events using relevant News and feedbacks.

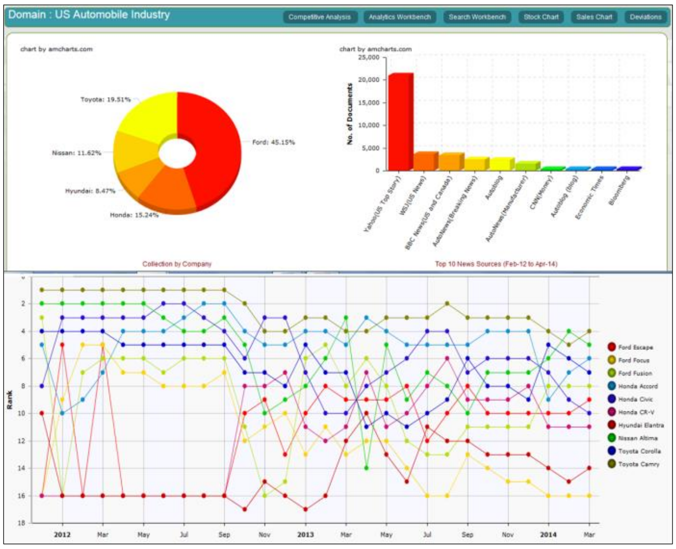


Figure 7 System snapshot

A. Case Study

For this case study we selected the model Nissan Altima whose monthly ranks for the period January 2012 to March 2013 are shown in Figure 8. This figure also high-lights two interesting deviation events identified by the system. For each deviation event, the system also shows the most relevant News articles retrieved from the News repository using the algorithm present in Section VII. It may be noted that, these are only two representative articles. There were many articles retrieved for each deviation-event which are ranked by decreasing order of impact computed as presented in section V.

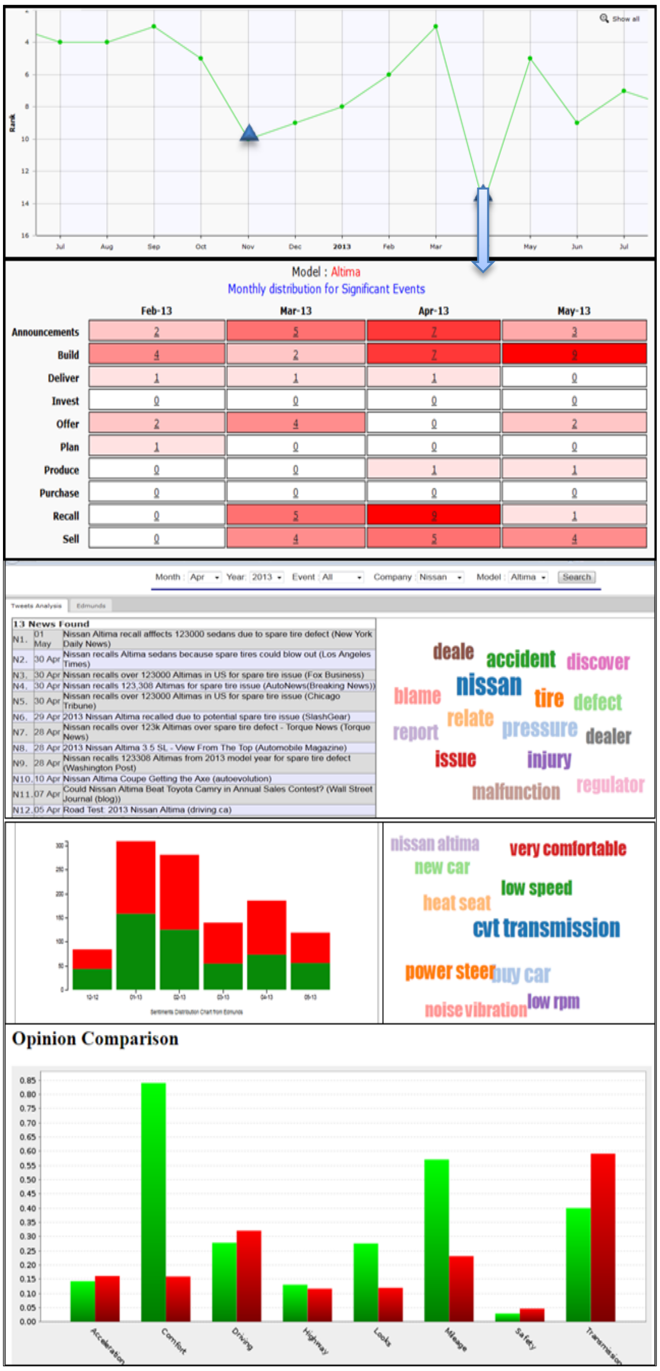
The April 2013 event is a negative one showing a huge fall in rank of this car. The Figure 8(b) shows a heat-map of different types of News events that were found to be significant around the period of interest. Figure 8(c) presents a list of News titles retrieved for the period. The most frequent event is the announcement of a recall of the model over a spare-tire issue. The tag-cloud on the right of figure c shows the prominently occurring phrases in these News articles. Figure 8(d) presents customer sentiments around this car and its components and features aggregated over the period of January 2013 to April 2013. While it is observed from the top left corner of figure (d), that positive and negative sentiments around the car are almost equal, the top right portion shows the frequently occurring features and attributes in these feedbacks. The bottom portion shows the opinion scores pertaining to the different attributes. It is clear that while customers had a positive feedback about comfort offered by the model, there were concerns about the "CVT Transmission, driving and safety".

Figure 9 plots normalized sentiment scores computed from customer feedbacks in Edmunds superimposed with the rank of this model computed from NADA data. It is interesting to observe the rise in negative sentiments from November 2012. The negativity continues till August 2013. When sentiment scores are correlated with the rank data, it is found that the correlation values are 0.07 and 0.44 with positive and negative sentiments respectively. This matches with known knowledge that customers take the pain to give negative feedbacks more often than positive feedbacks. However, the more important and interesting fact for an analyst is the discovery and confirmation that social media sentiments could serve as an indicators for future sales volumes.

IX. CONCLUSION

In this paper, we have presented a framework that facilitates integration and analysis of multi-structured data acquired from multiple sources specifically to enrich business intelligence. The paper also presents methods for integrating structured enterprise data like sales figures or stock-index reports with Open Source News articles and consumer-generated social-media content. It was also shown that data integration from multiple domain aid the analysis process by automatically capturing and correlating significant events across different data domains.

Event-based analytics paves the way for performing causal analysis. Events here present significant incidents reported in News collections. Events can be interpreted as causes or effects for data values recorded in structured data. Our future work includes extension of the framework for automatic assessment of impact of different categories of events over structured data. We envisage the framework to develop into an automated causal analytics framework with assimilation of more data and human inputs, which can be stored. Inference mechanisms can be learnt from the inputs using machine learning. The framework is being extended to capture and learn from human inputs. Thus co-occurrence relationships identified by the system initially from mere associations can be further refined to cause-effect relationships based on human inputs.



Model : Altima

Monthly distribution for Significant Events

	Feb-13	Mar-13	Apr-13	May-13
Announcements	2	5	7	3
Build	4	2	7	9
Deliver	1	1	1	0
Invest	0	0	0	0
Offer	2	4	0	2
Plan	1	0	0	0
Produce	0	0	1	1
Purchase	0	0	0	0
Recall	0	5	9	1
Sell	0	4	5	4

Month : Apr Year: 2013 Event : All Company : Nissan Model : Altima Search

13 News Found		
N1.	01 May	Nissan Altima recall afffects 123000 sedans due to spare tire defect (New York Daily News)
N2.	30 Apr	Nissan recalls Altima sedans because spare tires could blow out (Los Angeles Times)
N3.	30 Apr	Nissan recalls over 123000 Altimas in US for spare tire issue (Fox Business)
N4.	30 Apr	Nissan recalls 123,308 Altimas for spare tire issue (AutoNews(Breaking News))
N5.	30 Apr	Nissan recalls over 123000 Altimas in US for spare tire issue (Chicago Tribune)
N6.	29 Apr	2013 Nissan Altima recalled due to potential spare tire issue (SlashGear)
N7.	28 Apr	Nissan recalls over 123k Altimas over spare tire defect - Torque News (Torque News)
N8.	28 Apr	2013 Nissan Altima 3.5 SL - View From The Top (Automobile Magazine)
N9.	28 Apr	Nissan recalls 123308 Altimas from 2013 model year for spare tire defect (Washington Post)
N10.	10 Apr	Nissan Altima Coupe Getting the Axe (autoevolution)
N11.	07 Apr	Could Nissan Altima Beat Toyota Camry in Annual Sales Contest? (Wall Street Journal (blog))
N12.	05 Apr	Road Test: 2013 Nissan Altima (driving.ca)

Figure 8 (a-d)

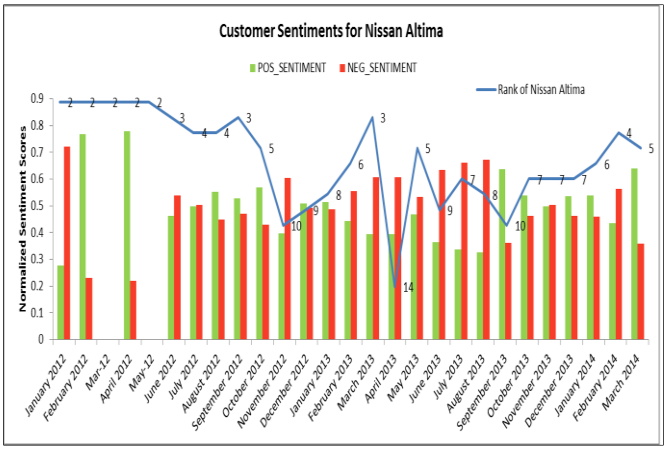


Figure 9 Customer Sentiments for Nissan Altima

REFERENCES

[1] Chinatsu Aone and Mila Ramos-Santacruz, “REES: A Large-Scale Relation and Event Extraction System”, in ANLC ’00 Proceedings of the Sixth Conference on Applied Natural Language Processing, 2000, Pages 76-83.

[2] J. Piskorski, H. Tanev, M. Atkinson and E. Van Der Goot, “Cluster-Centric Approach to News Event Extraction”, in Proceedings of the 2008 conference on New Trends in Multimedia and Network Information Systems, 2008, pages 276-290.

[3] F. Hogenboom, F. Frasincar, U. Kaymak and F. de Jong, “An Overview of Event Extraction from Text” in Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE'11).

[4] A. Das Sarma, A. Jain and C. Yu, “Dynamic Relationship and Event Discovery”, in WSDM, 2011.

[5] R. Reichart and R. Barzilay, “Multi Event Extraction Guided by Global Constraints,” in Proc. 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, Canada, 2012, pages 70–79.

[6] A. Ritter, Mausam, O. Etzioni and S. Clerk, “Open Domain Event Extraction from Twitter”, in KDD 2012, Proc. Of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, Pages 1104-1112.

[7] Lipika Dey, Ishan Verma, Arpit Khurdiya and Sameera Bharadwaja, A Framework to Integrate Unstructured and Structured Data for Enterprise Analytics, FUSION 2013, Istanbul, Turkey.

[8] Yiming Yang, Jaime Carbonell, Ralf Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu. Learning approaches for detecting and tracking news events. In IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, volume 14 (4), pages 32–43, 1999.

[9] Bhide, M. A., Gupta, A., Gupta, R., Roy, P., Mohania, M. K., & Ichhaporia, Z. (2007, June). Liptus: Associating structured and unstructured information in a banking environment. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (pp. 915-924). ACM.

[10] Bhide, M., Chakravarthy, V., Gupta, A., Gupta, H., Mohania, M., Puniyani, K., ... & Sengar, V. (2008, April). Enhanced business intelligence using EROCS. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on* (pp. 1616-1619). IEEE.

[11] Dey, L., & Haque, S. M. (2009). Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition*, *12*(3), 205–226.